

# Weekly Report

Pingping Shang

2013.10.14~2013.10.20

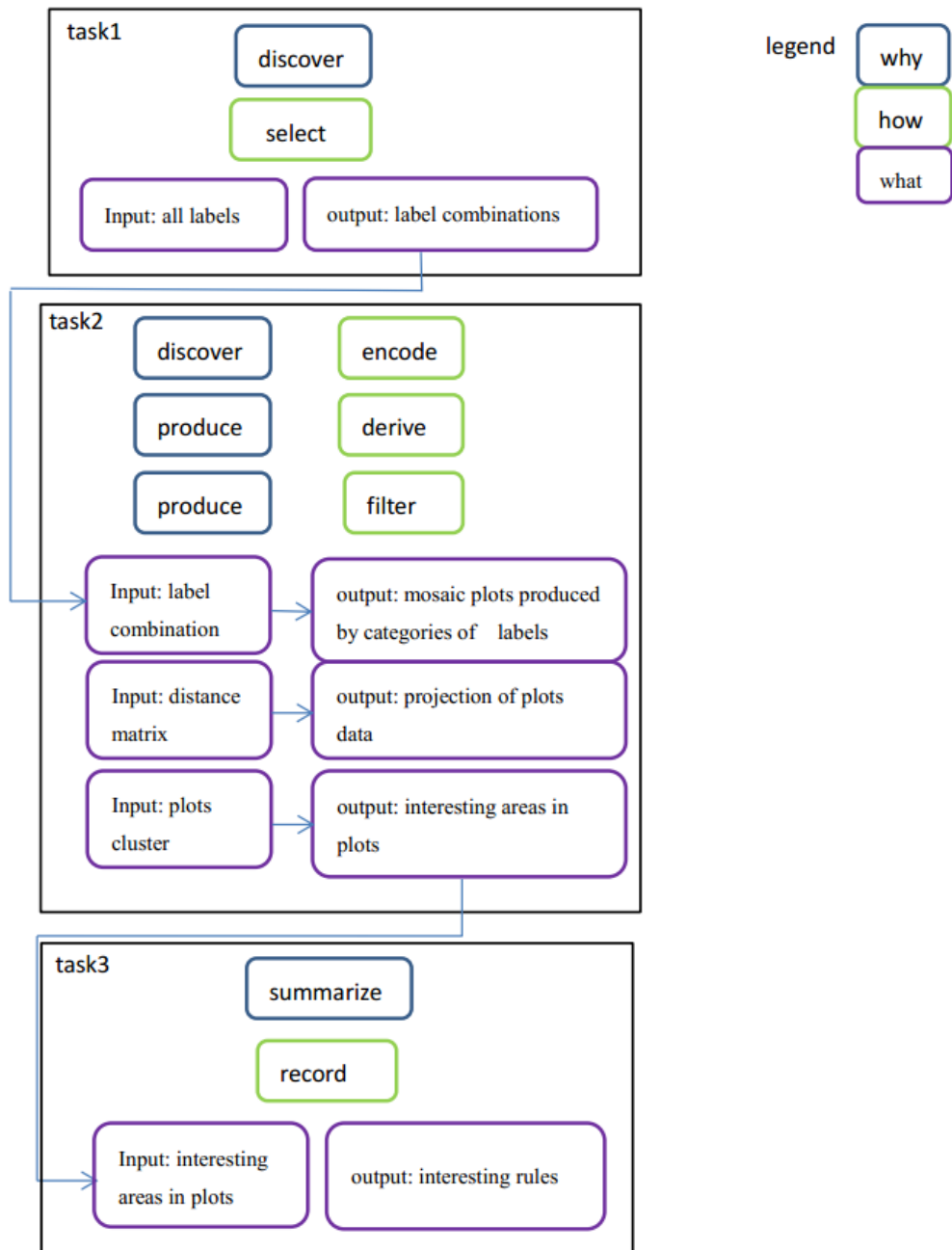
## 本周工作

- 1 试验 SOM 对各个标签的分类结果，年龄、性别、购买等级等标签都可以较明显被分开。但是标签组合得到的 SOM 分类结果就不太明显了。
- 2 用 CCA 分析标签间、行为间的相关性。我们的思路：
  - 1) 选定标签，按之前的方法将人群对应分好；
  - 2) 任两个人群间利用 CCA 计算相关系数，系数大的我们认为这两个人群有较大关联；此外，在这过程中得到最大相关系数的同时，得到了使得相关系数最大的行为名称，然后对这种行为计数，在人群对（pair）中出现次数最多的商品，我们认为它对不同人群的区分有较大贡献（该思路目前还没计算出好的结果，理论上也有待考证）

## 下周工作

做好 CCA 分析结果，若是结果还可行，再继续探索将这种分析结果怎么和我们前面做的工作结合起来。

## 一、“用户标签与用户行为探索”流程图



## 二、关联规则挖掘综述

### 1 概述

关联规则挖掘的一个典型例子是购物篮分析。关联规则研究有助于发现交易数据库中不同商品（项）之间的联系，找出顾客购买行为模式，如购买了某一商品对购买其他商品的影响。分析结果可以应用于商品货架布局、货存安排以及根据购买模式对用户进行分类。

围绕关联规则的研究主要包括：对原有的算法进行优化，如引入随机采样、并行的思想等以提高算法挖掘规则的效率；对关联规则的应用进行推广

此外，也有一些工作注重于对挖掘到的模式的价值进行评估。

## 2 算法描述

设  $I = \{i_1, i_2, \dots, i_m\}$  是项集，其中  $i_k (k=1, 2, \dots, m)$  可以假设是购物篮中的物品。设任务相关的数据  $D$  是事务集，其中每个事务  $T$  是项集，使得  $T \subseteq I$ 。

关联规则是如下形式的逻辑蕴涵： $A \Rightarrow B$ ,  $A \subset I$ ,  $B \subset I$ , 且  $A \cap B = \Phi$ 。

关联规则具有如下两个重要的属性：

1) 支持度： $P(A \cup B)$ ，即  $A$  和  $B$  这两个项集在事务集  $D$  中同时出现的概率。

2) 置信度： $P(B / A)$ ，即在出现项集  $A$  的事务集  $D$  中，项集  $B$  也同时出现的概率。

同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。给定一个事务集  $D$ ，挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则，也就是产生强规则的问题。

## 3 算法综述

### 3.1 经典的频集算法

核心是基于两阶段频集思想的递推算法。首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。然后由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。总体性能由第一步决定。

### 3.2 改进的频集算法

散列、事务压缩、杂凑、划分等，都是为了提高寻找频繁集的效率而提出的改进算法，由于跟我们的项目关系不是太大，在此不再细说。

## 4 多层关联规则挖掘

多层对应多个抽象层。对于很多的应用来说，由于数据分布的分散性，所以很难在数据最细节的层次上发现一些强关联规则。当我们引入概念层次后，就可以在较高的层次上进行挖掘。虽然较高层次上得出的规则可能是更普通的信息，但是对于一个用户来说是普通的信息，对于另一个用户却未必如此。所以数据挖掘应该提供这样一种在多个层次上进行挖掘的功能。

多层关联规则的分类：根据规则中涉及到的层次，多层关联规则可以分为同层关联规则和层间关联规则。

## 4 多维关联规则挖掘

多维对应多个维或谓词。对于多维数据库而言，除维内的关联规则外，还有一类多维的关联规则。例如：

年龄(X, “20...30”) 职业(X, “学生”) ==> 购买(X, “笔记本电脑”)

在这里我们就涉及到三个维上的数据：年龄、职业、购买。

根据是否允许同一个维重复出现，可以又细分为维间的关联规则（不允许维重复出现）和混合维关联规则（允许维在规则的左右同时出现）。

比如年龄(X, “20...30”) 购买(X, “笔记本电脑”) ==> 购买(X, “打印机”)就是混合维关联规则。

## 5 挖掘稀有模式和负模式

有时令人感兴趣的不是频繁模式，而是发现稀有的，或发现反映项之间的负相关模式。

稀有模式，由于大多数频繁集的出现频度通常都低于最小支持阈值，因此实践中都会增加一些额外条件。

如果项集 X 和项集 Y 都是频繁的，但很少一起出现，则它们两个是负相关的。

## 6 项目中可视化可能的落脚点

在挖掘的过程中，提供一种与用户进行交互的方法，将用户的领域知识结合在其中；生成结果的可视化。

结合我们的项目（用户标签和用户行为分析），主要有以下几方面考虑：

- 1) 我们想要找到的用户标签和用户行为间的关系，可看作多维关联规则挖掘，如前面混合多维的例子，若是直接采用关联规则挖掘的方法，应该也可以找到我们想要的，加入可视化的优势：将规则产生过程展示给用户？直观理解产生的规则在整个数据中所处的地位？（如我们的像素图中，只有一块是奇特的，其余都是随机形状）。
- 2) 标签选择相当于加入用户知识，只分析我们认为可能会有意义的标签组合？
- 3) 标签选定后，我们的目前的可视化只能算是统计结果的呈现，而且可视化过程也将所有事务遍历了一遍，未提高关联规则中寻找频繁项的效率。比如计算像素图中每一个像素值时，我们要遍历所有事务，记录它的绝对值。